Student: Riato Stefano
ID: 894041
Ca' Foscari - [ET7008] Lab of Information Systems and Analytics

Introduction

In the competitive realm of banking, leveraging data to forecast customer behaviors can significantly enhance the effectiveness of marketing strategies. This report focuses on using machine learning techniques to predict the likelihood of clients at a Portuguese banking institution subscribing to term deposits, based on data from direct marketing campaigns. Such predictive analytics aim to refine customer engagement approaches, potentially boosting subscription rates and optimizing the allocation of marketing resources.

Data Description

The analysis utilizes the "Bank Marketing Dataset" from the UCI Machine Learning Repository, specifically the `bank-additional-full.csv` file, which includes comprehensive data on 41,188 clients and spans from May 2008 to November 2010. The dataset encapsulates various attributes:

- Client Data: Includes age, job, marital status, education, default history, average yearly balance, housing, and personal loans.
- Campaign Data: Details on the contact type, last contact timing, call duration, campaign contact count, days since last contact from a previous campaign, and previous campaign outcomes.

Goal of the Project

The project's primary aim is to develop a machine learning model that predicts whether a client will subscribe to a term deposit (label 'y'). This prediction is crucial for tailoring the bank's marketing strategies more efficiently, ensuring better resource allocation and enhancing the success rates of these campaigns.

Data Cleaning and Exploratory Data Analysis (EDA)

Data Cleaning

The foundation of a successful analysis in data science starts with thorough data cleaning. For our dataset consisting of 45,211 records and 17 attributes, initial inspections revealed a well-compiled data set with no missing values across the board, which is relatively rare in real-world data scenarios. However, the 'pdays' attribute, indicating days since the last contact from a previous campaign, mostly registered as '-1', denoting no prior contact. Given that over 80% of the 'pdays' entries were '-1', this attribute was removed to streamline the dataset and focus on more impactful variables.

The dataset's integrity was further refined by addressing 'unknown' entries in significant categorical variables such as 'job', 'education', 'contact', and 'poutcome'. Each 'unknown' entry potentially dilutes the predictive strength of our models. For 'education' and 'job', where 'unknown' statuses were comparatively low, these entries were removed to preserve the robustness of those categories. For 'poutcome' and 'contact', which had a high proportion of 'unknowns', the entire attributes were dropped to avoid skewing our analysis.

Exploratory Data Analysis (EDA)

Post-cleaning, we delved into the Exploratory Data Analysis to unearth any underlying patterns or insights that could aid in building robust predictive models.

Target Variable Analysis:
The target variable 'y', which indicates whether a client subscribes to a term deposit, was significantly skewed. With around 88.3% of clients not subscribing, the data exhibited a pronounced imbalance that could potentially bias predictive modeling towards the majority class. This aspect called for specialized techniques in handling imbalanced data to ensure both classes are predicted with high accuracy.

Categorical Variable Insights:
Exploration of categorical variables revealed distinct trends; for instance, clients with management jobs or a tertiary education level showed a slightly higher propensity for subscribing to term deposits. Conversely, those with blue-collar jobs or basic education were less likely to subscribe. The month of contact also seemed to play a role, with months like May experiencing higher contact rates but not necessarily higher subscription rates, suggesting a possible fatigue effect.

Numerical Data Distribution:
Numerical attributes such as age, balance, and call duration were analyzed for distribution patterns. The age and balance variables were right-skewed, indicating that a majority of clients were younger and with lower yearly balances. The call duration showed a wide range with a peak in shorter calls, hinting that shorter, possibly more efficient calls could be more frequent but not necessarily more effective in achieving subscriptions.

Pre-processing and Feature Engineering

Data Pre-processing

Initial investigations into the dataset's structure revealed eight nominal features without any ordinal or continuous attributes, pointing to a dataset predominantly categorical in nature. These features include job type, marital status, education level, default history, housing and personal loan status, the month of last contact, and the target variable y (subscription outcome). The absence of ordinal features simplifies our approach, allowing us to primarily focus on one-hot encoding to transform these nominal variables for better analysis and prediction.

One-hot encoding was applied to convert categorical variables into a format that could be more easily interpreted by machine learning algorithms.

Feature Engineering

A critical step in feature engineering was the analysis of feature correlations to identify any redundant pairs of features that might lead to multicollinearity, where highly correlated features can distort the importance of variables in some models. Our correlation analysis revealed several pairs of binary variables created from one-hot encoding that were perfectly correlated with each other, such as 'loan_yes' and 'loan_no'. Additionally, some features like 'marital_married' and 'marital_single' showed a significant correlation, suggesting an overlap in the information they provide.

Based on the correlation threshold of 0.7, we identified and removed features that were redundant, thereby simplifying the model's complexity without sacrificing predictive power. This step helps in reducing overfitting and improves the generalization of the model on unseen data.

Scaling Numerical Features

To further refine our model, numerical features were scaled using StandardScaler, a common preprocessing technique that standardizes features by removing the mean and scaling to unit variance. This technique transforms the data such that its distribution will have a mean value 0 and standard deviation of 1, ensuring that each feature contributes equally to the distance computations in the model, an essential aspect especially for distance-based algorithms like KNN. By standardizing the data, we not only improve the convergence during training but also enhance the model's performance by treating all features equally, thus preventing any single feature with a higher range from dominating the predictive process.

Model Training and Evaluation

Data Splitting

The dataset was meticulously divided into three distinct sets: training, validation, and test. This separation facilitates an unbiased evaluation of the models, ensuring that they are trained on one subset of the data and validated and tested on completely independent subsets. Specifically, 80% of the data was used for training, with the remaining 20% for testing. Further, 20% of the training set was reserved for validation purposes. This structured approach aids in mitigating overfitting and validating the model's effectiveness before final testing.

Model Training

K-Nearest Neighbors (KNN)

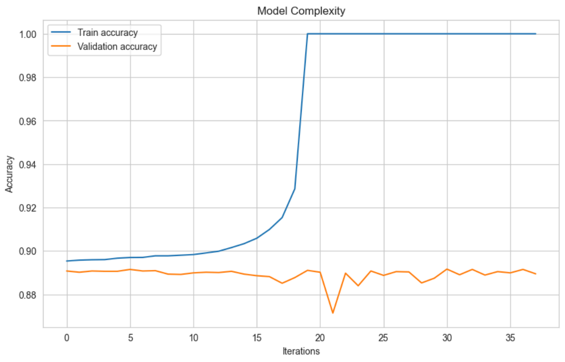
The KNN algorithm was implemented with initial hyperparameters set at three neighbors and uniform weights. The model's performance was assessed using the F1 score, a critical metric given the dataset's imbalance. The validation F1 score and accuracy were calculated to ensure the model's capability to generalize beyond the training data.

Random Forest

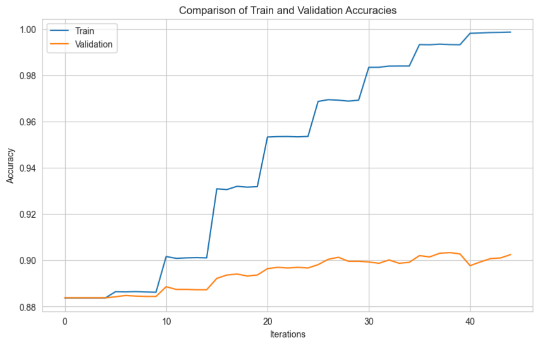
Random Forest was employed due to its efficacy in handling large datasets and capability to model non-linear relationships. The model was initially configured with 100 trees. Subsequent tuning involved adjusting the number of trees and the maximum depth to strike a balance between learning detailed data patterns and avoiding overfitting.

Hyperparameter Tuning

A grid search was conducted over a range of parameters to fine-tune the KNN model and also the Random Forest. This process was instrumental in identifying the optimal settings that maximize validation accuracy, thereby enhancing the model's performance.



KNN



Random Forest

The graphs illustrate the training and validation accuracy during the hyperparameter tuning process, highlighting the model's complexity and its impact on performance. As the iterations increase, the training accuracy sharply rises, reaching near-perfect accuracy. This indicates that the model becomes highly tuned to the training data, suggesting a possible overfitting scenario where the model learns the training data's specific characteristics too well, including noise and outliers.

Conversely, the validation accuracy remains relatively constant and significantly lower than the training accuracy throughout the iterations. This disparity between training and validation performance is a classic indicator of overfitting. The model, while performing exceptionally well on the training dataset, fails to generalize effectively to unseen data represented by the validation set.

Gaussian Naive Bayes

As a probabilistic classifier, Gaussian Naive Bayes was chosen for its simplicity and efficiency in handling binary classification tasks. It was evaluated based on its accuracy and F1 score on the test set to confirm its predictive power and reliability.

Model Evaluation

Validation and Test Results

Each model was rigorously evaluated on both validation and test datasets. Performance metrics such as accuracy and F1 score were computed to assess each model's effectiveness in correctly predicting the outcomes.

Model	Test Set scores	F1 score on Test set
KNN	0.894	0.628
Random Forest	0.901	0.706
Gaussian Naive Bayes	0.864	0.653

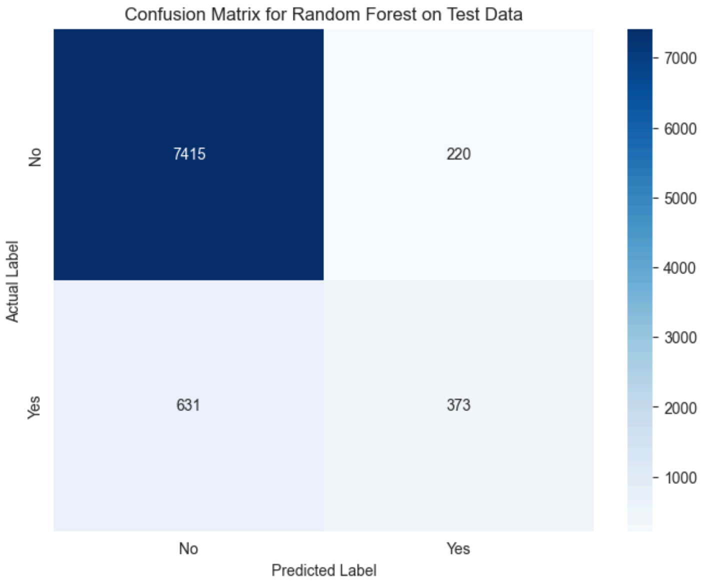
The Random Forest model exhibited superior performance, registering the highest accuracy at 0.901493 and an F1 score of 0.706428. This indicates not only a high rate of correct predictions but also an effective balance between precision and recall, crucial for handling the imbalanced nature of our dataset.

On the other hand, the KNN model, while achieving a commendable accuracy of 0.894548, had a lower F1 score of 0.628701. This suggests that while KNN is relatively adept at identifying the correct outcomes, it may not equally capture the nuances of both positive and negative classes. Gaussian Naive Bayes, despite having the lowest accuracy of 0.864105, maintained a moderately high F1 score of 0.653839, underscoring its potential usefulness in specific scenarios where true positive and true negative predictions are balanced against computational efficiency.

The choice of the model would therefore hinge on specific project requirements: Random Forest is ideal for maximizing prediction accuracy and managing class imbalance effectively, making it suitable for our current objective of optimizing marketing strategies. However, for tasks requiring faster processing times at a slight compromise on accuracy, Gaussian Naive Bayes might be considered. KNN could be useful in applications where the proximity of data points significantly influences the outcome. The choice of the best model for our goal is Random Forest.

Confusion Matrix Analysis

The confusion matrix highlights the Random Forest model's strengths and weaknesses. While it is quite accurate overall, it tends to miss a substantial portion of the clients who do subscribe (low recall), which could mean missed opportunities in a real-world scenario. The model is conservative, minimizing the risk of false positives but at the cost of missing true positives. These insights are crucial for refining the model further or adjusting the threshold to balance precision and recall according to the bank's marketing strategy objectives.



Conclusion

This project demonstrates the Random Forest model's superiority in predicting term deposit subscriptions due to its high accuracy and F1 score, effectively handling the class imbalance in the dataset. Its robust performance confirms its suitability for optimizing the bank's marketing strategies, making it the preferred model for enhancing customer engagement and increasing subscription rates. This choice underscores the importance of leveraging advanced analytics to drive strategic decisions and improve business outcomes in the banking sector.